

Statistical approaches in weed research: choosing wisely

Métodos estatísticos em pesquisa com plantas daninhas: escolhendo adequadamente

Germani Concenção¹, Andre Andres², Fabio Schreiber³, Ananda Scherner⁴, Joao Pedro Behenck⁵

Abstract - Statistical concepts and methods play an important role in the society, and statistical data analysis require considerable human labor and knowledge. From one side, computers and statistical softwares allow almost anyone to run free on statistical methods, but on the other side any researcher, professor, student or professional, even lacking on basic statistical knowledge to test their data, may use these softwares, often producing biased statistical analyses. The objective of this review is to demonstrate how the choice for statistical methods in weed science may create a bias in the interpretation of herbicide efficiency, and impact herbicide recommendations. We propose minor changes to the ordinary approach to help avoiding data misinterpretation and unintentional erroneous herbicide recommendations. The problems discussed throughout the review are illustrated with real field experimental data. Great part of the results of studies involving herbicide efficacy seems to be based on underpowered experiments and prone to output distorted information. Flawed choices of statistical methods, specially the p-value based statistics (ANOVA and *post-hoc* tests), can pave the way for mistaken conclusions even in properly conducted experiments in weed research. It is proposed the use of confidence intervals for both qualitative and quantitative data analysis, coupled to an appropriate number of samplings (“n”).

Keywords: ANOVA, multiple mean comparison, regression analysis, statistical bias, confidence intervals.

Resumo - Conceitos e métodos estatísticos possuem papel fundamental para a sociedade, e a análise estatística de dados demanda considerável esforço e conhecimento humano. Por um lado, computadores e softwares estatísticos permitem que virtualmente qualquer pessoa possa escolher e executar testes estatísticos, mas por outro lado qualquer pesquisador, professor, estudante ou profissional pode utilizar estes softwares, mesmo os que não possuem conhecimentos estatísticos básicos para testar seus dados, produzindo com frequência análises estatísticas com algum tipo de incorreção. Objetiva-se com a presente revisão demonstrar como a escolha do método estatístico na ciência das plantas daninhas pode criar um viés na interpretação da eficiência de herbicidas, e impactar

Received: April 13, 2017. Accepted: June 20, 2017.

¹ Empresa Brasileira de Pesquisa Agropecuária – EMBRAPA, Sistemas Integrados de Produção, Embrapa Clima Temperado, Rodovia BR-392, Km 78, 9º Distrito, Monte Bonito, CP 403, CEP 96010-971, Pelotas, RS, Brazil. E-mail: germani.concencao@embrapa.br

² Empresa Brasileira de Pesquisa Agropecuária – EMBRAPA, Manejo Sustentável de Plantas Daninhas, Embrapa Clima Temperado, Pelotas, RS, Brazil. E-mail: andre.andres@embrapa.br

³ Universidade Federal de Pelotas – UFPEL, Plantas Daninhas, Capão do Leão, RS, Brazil. E-mail: schreiberbr@gmail.com

⁴ Aarhus University, Plant Science, Aarhus, Denmark. E-mail: anandascherner@hotmail.com

⁵ Universidade Federal de Pelotas – UFPEL, Faculdade de Agronomia, Capão do Leão, RS, Brazil. E-mail: joaobehenck@gmail.com

sua recomendação. Os problemas discutidos são ilustrados com base em dados de experimentos reais de campo. Propomos pequenas alterações na forma atual de análise de dados para auxiliar a reduzir a má interpretação de dados e a equivocada recomendação de herbicidas com base em sua eficiência em experimentos. Grande parte dos resultados de estudos com herbicidas parece estar embasado em experimentos sem o devido poder estatístico, logo sujeitos a fornecer informações com viés. A escolha por métodos estatísticos falhos, especialmente os baseados no valor-p (ANOVA e testes *post-hoc*), podem estar levando a conclusões equivocadas mesmo em experimentos com herbicidas corretamente conduzidos. Propõe-se o uso de intervalos de confiança para análise de dados qualitativos e quantitativos, junto ao adequado número de amostragens (“n”).

Palavras-chave: ANOVA, comparação múltipla de médias, análise de regressão, viés estatístico, intervalos de confiança.

Introducing an old problem

Statistical concepts and methods play an important role in the society, since most decisions on science and technology are guided by its significances. From toys to aircrafts, engineering to social sciences and policy making, statistics is a decision making element (Silveira Junior et al., 1989). The main factor establishing statistics as such an important tool is the ability to infer about traits of large populations by collecting and processing data of small population samples, which cannot be studied as a whole (Silveira Junior et al., 1989; Steel and Torrie, 1980).

Traditional statistics is based on books full of equations and complex mathematical terms which need to be understood and calculated; thus, analysis of statistical data requires considerable human labor and knowledge (Steel and Torrie, 1980). Many research institutions used to have full time statisticians supporting exclusively experimental data analysis. The advent of informatics made the work of processing experimental data easier and faster due to the broad diversity of statistical softwares available (Paternelli and Mello, 2011) and, therefore, the dependency on statisticians has been decreased.

On the bright side, computers and statistical softwares allow researchers to run free on statistical methods (Paternelli and Mello, 2011). However, these tools are available for any researcher, professor, student or professional,

that may lack on basic statistical knowledge to test their data, consequently inducing to wrong assumptions (Reinhard, 2015). The statistical software alone is completely dumb and useless, therefore, to produce statistical analyses with substantial meaning these should be based on the correct hypothesis and in reliable datasets (Steel and Torrie, 1980; Silveira Junior et al., 1989).

Great part of the agronomy students which specialize in weed science quickly learn the following three elementary rules of current statistics (Steel and Torrie, 1980):

- Traditional experimental statistics work with an accepted error rate of 5%;
- F-test is the almighty one, the God of the experimental statistics;
- Mean comparisons from a herbicide test should be accomplished by using Tukey's test because Duncan's MRT has greater risk of outputting false positive (Type I) errors.

However, there are other issues that should be considered when choosing the experimental statistics approach (Huff, 1954; Reinhard, 2015). In fact, the rules presented above have been the focus of extensive discussions and manuscripts can be easily rejected by reviewers based on a simple author's choice of a post-hoc statistical test.

Weed scientists should deepen their statistical knowledge to better explore experimental data. Sometimes the information we aim for is implicit

in the dataset, but the statistical procedures chosen are not able to show it. Furthermore, the use of correct statistics would also avoid the introduction of the “bad pharma” effect (Goldacre, 2012) in the weed science, resulting in the recommendation of herbicides which are not suitable. Goldacre’s book (2012) leaves an important message about the correct application of statistics and ethics in research, interpretation of results and publishing, though with a rather coercive and particular writing style. These advices are important because researchers often produce exaggerated results, even with no malicious intent, that strongly favor their hypothesis (Reinhard, 2015).

The objective of this review is to demonstrate how the choice of statistical method to analyze experimental data in weed science may create a bias in the interpretation of herbicide efficiency, and impact herbicide recommendations. In addition, we propose here some minor changes to the ordinary approach that weed scientists adopt to analyze and interpret data, to help avoiding data misinterpretation and unintentional erroneous herbicide recommendations.

The reader should note, however, that the present review was not written by statisticians; neither there was any intention to go deeper in an issue which is not the specialty of the authors, nor to focus in statistical formulas and concepts. The information is supplied in practical terms and illustrated as much as possible with real experimental data.

The underpowered statistics

Experimental statistics usually applied to the weed science are parametric approaches based on a threshold value (Steel and Torrie, 1980), which is the significance level of the F-test. If the threshold level of significance is reached (e.g., supposing $p \leq 0.05$) a post-hoc test should be performed, which may be either a multiple mean comparison procedure (for qualitative data) or regression analysis (for quantitative data). The test choice for mean comparison or alternatively for the

class of the regression to be adjusted to the data are the topics of extended statistical discussions (Reinhard, 2015).

Statistical methods heavily based on the p-value, as weed science researchers are familiar with, are often criticized by statisticians (Cleveland, 1979; Cumming et al., 2004; Reinhard, 2015). They consider this approach as not efficient or reliable, mainly by not supplying any clue about the size of the treatment effects which are studied and, for providing only partial answers to researchers. The main problem with p-value based statistics will not be discussed but illustrated with data analysis from a real field experiment.

The experiment which produced the results used in the analyses at Table 1 was installed under field conditions in a randomized blocks design with nine replications. Each plot comprised three rows of 15 m spaced in 4 m, where *Jatropha* plants were planted spaced in 1 m. Thus, each replicate of the given experiment was composed by 45 *Jatropha curcas* plants that were 2-year old. Six residual herbicides were applied to the inter-rows in August 2013 aiming to evaluate both the weed control levels and crop toxicity. Crop productivity was evaluated in early 2014 by harvesting all plants in the plot. Moreover, the experiment was surrounded by *Jatropha* plants to avoid the “border effect”. The dataset from this experiment was analyzed by the F-test at 5% probability in four different ways as follow:

- I- Three replications, completely randomized design (CRD);
- II- Nine replications, CRD;
- III- Three replications, randomized blocks design (RBD);
- IV- Nine replications, RBD.

Mean comparison was accomplished by Duncan’s MRT test at the same probability level. Results from the four analyses are given in Table 1.

Table 1. Seed productivity of 2-year old *Jathropa curcas* plants, after application of six residual herbicides in the fall period for long-term weed control. Embrapa Western Agriculture, Dourados-MS, Brazil, 2013-2014.

CRD, first 3 replications		CRD, 9 replications	
F-test: 163.49 / p-value: < 0.001 / SW _{p-value} : 0.37 ^{YES}		F-test: 7.25 / p-value: < 0.001 / SW _{p-value} : < 0.01 ^{NO}	
Treatment	Productivity (kg ha ⁻¹)	Treatment	Productivity (kg ha ⁻¹)
2-Hoeing	1250.3 a	6-Herbicide D	1021.2 a
6-Herbicide D	1227.6 a	2-Hoeing	1005.0 a
8-Herbicide F	1151.6 b	8-Herbicide F	973.5 a
4-Herbicide B	1004.3 c	3-Herbicide A	824.0 a
3-Herbicide A	971.3 c	4-Herbicide B	754.0 ab
5-Herbicide C	692.3 d	5-Herbicide C	558.6 bc
1-Uncontrolled	573.3 e	1-Uncontrolled	554.1 bc
7-Herbicide E	548.3 e	7-Herbicide E	424.3 c

RBD, first 3 replications		RBD, 9 replications	
F-test: 1225.6 / p-value: < 0.001 / SW _{p-value} : 0.96 ^{YES}		F-test: 67.6 / p-value: < 0.001 / SW _{p-value} : 0.07 ^{YES}	
Treatment	Productivity (kg ha ⁻¹)	Treatment	Productivity (kg ha ⁻¹)
2-Hoeing	1250.3 a	6-Herbicide D	1021.2 a
6-Herbicide D	1227.6 a	2-Hoeing	1005.0 a
8-Herbicide F	1151.6 b	8-Herbicide F	973.5 a
4-Herbicide B	1004.3 c	3-Herbicide A	824.0 b
3-Herbicide A	971.3 d	4-Herbicide B	754.0 b
5-Herbicide C	692.3 e	5-Herbicide C	558.6 c
1-Uncontrolled	573.3 f	1-Uncontrolled	554.1 c
7-Herbicide E	548.3 g	7-Herbicide E	424.3 d

Original unpublished data from the authors. Herbicides were applied in August 2013, and harvest was accomplished in early 2014. SW = Shapiro-Wilk normality test (^{YES} = residuals can be considered normal; ^{NO} = residuals can NOT be considered normal). CRD = completely randomized design; RBD = randomized blocks design. Means followed by the same lowercase letter into each column, are not distinct according to the Duncan's MRT test at 5% probability.

Some problems can be pointed out from these analyses. First, by comparing CRD and RBD analyses with three replications (left side of Table 1), one may note that the same production levels result in different significances between treatments by simply changing the location of the plots into the field. If the experiment was installed in CRD, herbicides B and A were equivalent and the herbicide E was as bad as the uncontrolled treatment. In RBD, herbicide A had a worst performance than herbicide B. Moreover, herbicide E was even worse than in the previous scenario, doing probably nothing for weed control in *Jathropa* and causing more trouble to production than weed presence. Therefore, based on these results the owner of herbicide A would suggest that trials with this herbicide should be installed

in CRD, and thus the company may say that herbicide A is as safe to *Jathropa* as herbicide B, but cheaper. To avoid this, the experimental design should be always oriented only to correct restriction factors associated to the conditions into which the experiment will be installed. When comparing the analyses with 9 replications (right side of Table 1) in CRD and RBD, similar discrepancies can be observed.

By browsing into statistical textbooks while trying to understand how (or why) the results change considerably by only shifting the theoretical plot location into the field, one will find that this difference occurs because the *unobserved data* is different for each experimental design (Steel and Torrie, 1980; Reinhard, 2015), and because of this we have to focus on the errors instead

of the real data. A better explanation is that the experimental method based on p-values has a little of clairvoyance as it considers data that is “supposed to be there”, but no one sees and in fact they seem to do not exist.

Secondly, when considering the analyses with all replications (9), the data behaves distinctly between experimental designs. In CRD, the residuals are not even normally distributed (by Shapiro-Wilk test), while they are normal in RBD (Table 1). Therefore, in CRD data should be transformed (tortured in fact) prior to the F-test. To aggravate the case, researchers may have the habit of excluding replications because according to some test they are *outliers* but, an imbalanced analysis is probably not positive to an already delicate scenario (Donner and Koval, 1989).

Furthermore, there is the difference in treatment ranking and mean productivities between the experiments with three and nine replications. Since statistics is supposed to allow inferences about traits of large populations by processing data collected from small samples of the original population (Silveira Junior et al., 1989; Steel and Torrie, 1980), an adequate test was supposed to self-compensate and correctly rank the treatments independently of the number of replications. If this is not happening, it should be because the experiment may simply lack *power* enough to allow *any* affirmation. Results from unpowered experiments are kind of random and largely unreliable.

The power of herbicide experiments

One of the problems with the herbicide experiments that have been performed is the lack of *power* to make reliable assumptions if a given herbicide is superior (or not) to another one. The definition of power and how to test it experimentally is easily found in statistics textbooks and is beyond the scope of the present review. However, power definitions consider the Type I and Type II errors (Steel and Torrie, 1980).

Correctly powered experiments optimize the reality of treatment means. The power is proportional to the sample size used in the experiment, being also affected by the degree of variation among the replications and the significance level of the test (Steel and Torrie, 1980). For the experimental statistics, this means basically the number of replications (in fact, the total number of sampling points per treatment), and how carefully data was collected during plots evaluation (Steel and Torrie, 1980; Silveira Junior et al., 1989). Perry et al. (2003) report that for data which follow the normal distribution it can be easy to establish the power of standard tests. However, in the ecological and agricultural contexts, where count data frequently present an asymmetric distribution in relation to the mean, calculations of the power of an experiment may not be easy. The same authors carefully estimated the number of samples for a study to evaluate the impact of genetically modified crops with tolerance to herbicides in the United Kingdom. They concluded that 20 fields per crop per year (in other words, 20 replications or sampled points per treatment), over 3 years ($n = 60$), would provide enough power for the test.

Scientific journals may not require authors to repeat the experiment for at least two years, with data from a single experiment being enough for publication. Researchers often use between 3 to 5 replications per treatment within an experiment. These numbers of replications probably do not confer enough statistical power for herbicide-related trials, given the natural variation in the answers to the distinct herbicides or doses. Steel and Torrie (1980) reported that an experiment with 16 replications is about twice more efficient than one with four replications (standard deviations are at 2:1 ratio) in detecting treatment differences. Experiments with smaller number of replications depend mostly on the degrees of freedom for the error (residues) to be consistent. The power of the experiment, however, is so important that Reinhard (2015) reported that, if a

given experiment presents only about 50% power, and two compounds – let's say herbicides – are compared, in the first test the herbicide X will perform better than the herbicide Y; if we repeat the test, the opposite may be observed.

This also means that if a weed scientist tries to solve the problem of underpowered statistics by replicating an experiment with a given number of herbicide treatments (e.g. “6”) and replications (e.g. “4”) several times all throughout Brazil (e.g. 300 experiments), he is in fact installing a series of underpowered experiments, and results may be contrasting. It will be easy to find about 40 experiments, among the 300 ones, to support any affirmation, mainly if *p-value* is used for inferences. This may leave room for the undesired “bad-pharma” effect (Goldacre, 2012): if a result doesn't succeed (e.g., the herbicide we rely on is not the best one), the researcher has just to try again (Reinhard, 2015).

Limitations of the p-value for the weed science

As previously elucidated (Table 1), it is possible to obtain different p-values when considering different experimental designs (Reinhard, 2015), which makes p-values something “mystic” because distinct responses may come from the same dataset. In addition, statistics based on p-values are frequently misleading and confusing (Huff, 1954; Reinhard, 2015). One should remember that the usual statistics very often accepts a level of 5% error. In other words, about one experiment in every 20 is probably misleading and may generate what is called “statistical fallacy”. Figure 1 is adapted from the idea supplied by Reinhard (2015) to illustrate this phenomenon. The WeedScience.org list of herbicides lists 16 herbicides whose mechanism of action (MoA) is the inhibition of the enzyme ACCase, but this example assumes that there are 20 herbicides within this MoA.

According to Figure 1, the error chance is increased in studies where the F-test is run several times – like

in the evaluation of many variables from the same trial. To the error attributed to the “fallacy”, is added what is called the “base rate error” (Reinhard, 2015), which is associated to errors in experiment sampling and evaluation. For instance, while evaluating the control efficiency of herbicides on various weed species, the researcher is most prone to report the taller ones, while the prostrated weed species may sometimes pass unnoticed.

Another example of how statistics can ambush researchers which are unaware of the p-value limitations, is found in Ulguim (2016). This researcher investigated the difference in the light compensation point between two biotypes of *Euphorbia heterophylla*, one resistant and other susceptible to a given herbicide. The light compensation point for the susceptible one was $20 \mu\text{mol m}^{-2} \text{s}^{-1}$ while the resistant one presented $37 \mu\text{mol m}^{-2} \text{s}^{-1}$. The author concluded that the light compensation point was smaller for the susceptible one and that this biotype could probably perform better under reduced light regimes. This difference, however, represents only brief moments both at the beginning and ending of each day (Korczynski et al., 2002). In general, the larger the sample size in an experiment, more likely differences will be detected by a p-value test (Dahiru, 2008). In fact, it is possible to reach significance for about any difference in a p-value based test, by simply increasing the number of replications – by giving *power* to it.

The base rate error, added to the statistical fallacy and to the tendency to easily show treatment effects in properly powered experiments, shows that statistically significant results are false positives much more often than the $p < 0.05$ criterion for significance might suggest (Reinhard, 2015). Because of this, even properly applied statistics can't be completely trusted (Huff, 1954; Reinhard, 2015). In the medical area, there are detailed studies about the use of p-values. Chavalarias et al. (2016) report that among 1000 papers from the PubMed repository, almost all reported the use of p-values, which were significant (where are the

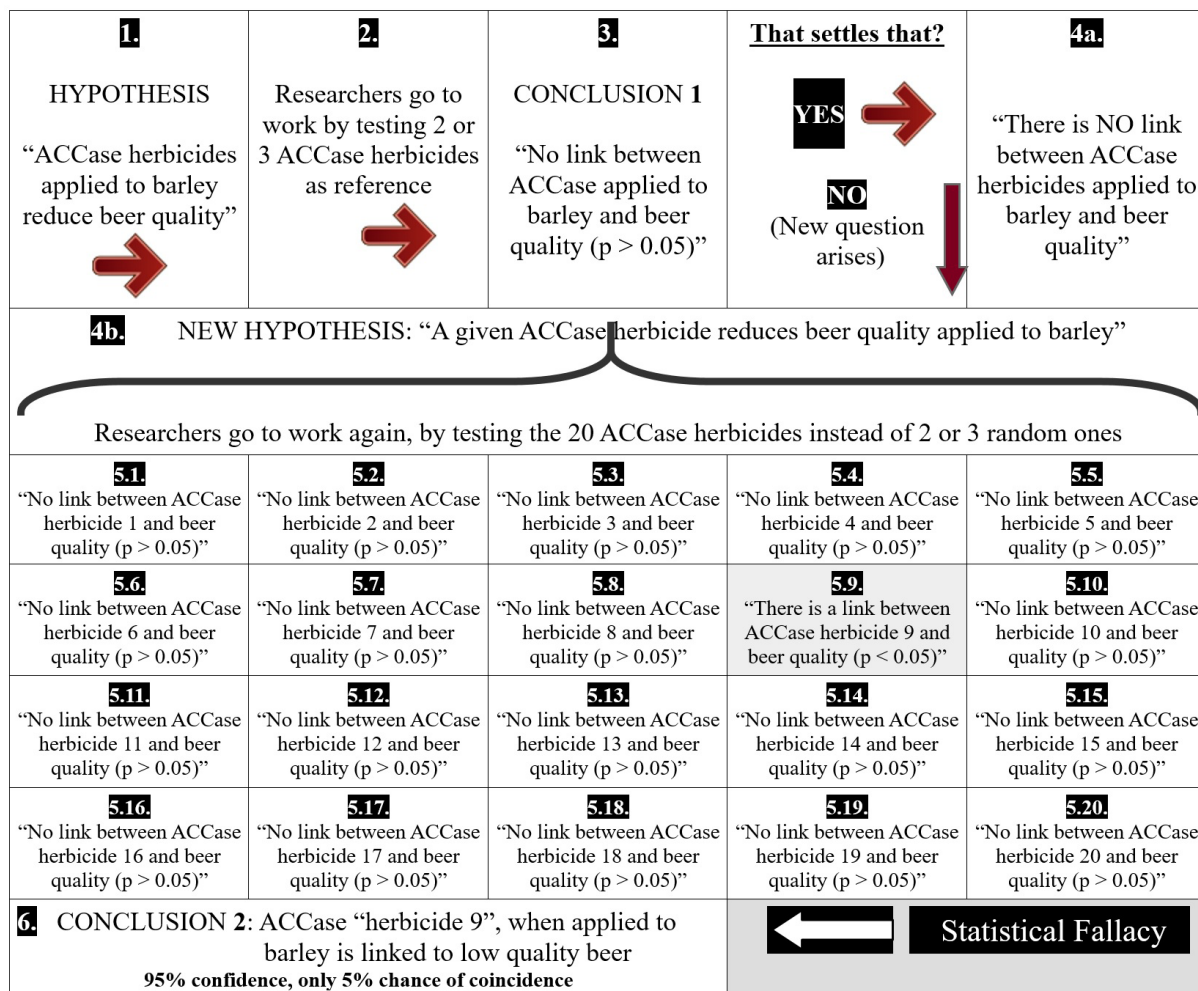


Figure 1. Schematic illustration of the statistical fallacy phenomenon with an accepted error rate of 5% ($5/100 = 1/20$). Source: adapted from the example supplied by Reinhard (2015). The depicted hypothesis is fictional and provided only as example.

not significant results?) and just a few included confidence intervals, Bayesian statistics, or other effect sizes. The authors recommended the inclusion of effect sizes and uncertainties metrics into the papers rather than reporting p-values.

For comparative purposes, the papers published in the journal *Planta Daninha* in 2016 (excluding literature reviews and other non-experimental documents) were screened for statistical methods (Figure 2). Although an indicative of publication quality, this is only a very rough estimation of the statistics used in the Brazilian weed science context, as it is based in only 73 papers from a

single Journal. A most extensive study could eventually show a different prospect.

From the 73 experimental papers screened, 37 (51%) were based only in p-values (Figure 2). About 85% described the experimental conditions and design, and more than 90% used replications. The mean of replication number per experiment was 4.2, with some experiments using 3 or 5, and few using 6 replications. Only one study used 10 replications. More than half the studies used only mean comparison or alternatively only regression analysis, with no kind of effect sizes; in fact, only about 32% of the studies presented confidence

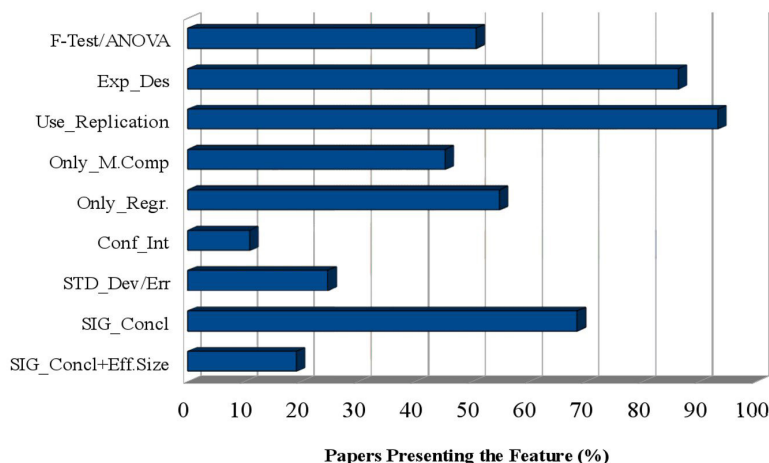


Figure 2. Adoption of statistical methods and/or data presentation in papers published at the Journal *Planta Daninha* (2016). Reviews and other non-experimental documents were removed from the analysis, being screened a total of 73 papers.

intervals, standard deviations or standard errors (Figure 2). About 68% of the papers used the word “significant” in the final considerations/conclusions (the last three paragraphs of the paper), while only 18% of them reported some kind of effect sizes together with the “significant” results in the conclusions.

The various limitations associated to the use of the p-value for the everyday statistics goes beyond the scope of this review, but Huff (1954) and Reinhard (2015) supply several case studies regarding the drawbacks of p-value applications. Some scientific journals encourage authors to make use of alternatives to the p-value based statistics, while others do not publish papers based on it, like the American Journal of Public Health did in the 1980's (Farland et al., 2016). Moreover, some authors believe that carefully conducted studies should be discussed by the scientific community, even if they lack statistical significance ($p > 0.05$).

Researchers should be aware that the word “significant” is so vague that it can be ineffective, though it is known that the addition of “highly” to it could stand for half a million dollars funding. However, “highly significant” does not classify experimental results as “most important” or “most reliable”; researchers should be careful

and avoid using flawed statistics, as well as to avoid over-value their findings based on these statistical methods, when reporting their data.

Alternatives to the p-value statistics

Some researchers advocate that p-values should be abandoned and statistical analyses based on confidence intervals or Bayesian methods. Another group believes the current statistics is just fine but not used correctly (Reinhard, 2015).

In the context of weed science, there is a need for experimental statistics and experimental designs that would enable researchers to obtain unbiased estimates of experimental errors, treatment means and significances (Steel and Torrie, 1980). The key point to be solved is that the experiment outputs should not be affected by the statistical design of the experiment, as occurs with statistics based on p-values (Table 1). This happens in the F-test/ANOVA most probably because we work with the errors for the analyses, instead of the real data. The experimental design should have the main role of controlling the experimental errors into the field / lab which are attributed to differential conditions in plot locations (Burns

and Dobson, 1981), but should not affect the final results and treatment ranking.

The less painful choice seems to be the implementation of confidence intervals to differentiate treatment effects (Steel and Torrie, 1980); this allows the use of experimental designs to control the external errors but makes treatment comparison independent of plot location or arrangement (Cumming et al., 2004). Rao et al. (2008) state that “effect sizes” should always be reported along with confidence intervals, being the “*minimum expectation*” for a reliable paper (APA, 2009).

Mean comparison with confidence intervals

In the p-value based statistics, researchers are familiar to use a F-test prior to mean comparison as a way to “protect” the mean comparison procedure (Steel and Torrie, 1980). This is completely unnecessary, and the called “post-hoc” tests (Duncan’s MRT, Tukey, LSD, Bonferroni, SNK and others) may be executed independently of a previous F-test, and even differences identified by these tests where the F-test indicated no difference among treatments, are valid (Hsu, 1996). Running ANOVA prior to mean comparison tests (“protecting” the post-hoc test) is more a tradition than a statistical requirement.

Following the example supplied at Table 1, the same data set was used to estimate treatment effects and effect sizes by using 95% confidence intervals (Table 2; Figure 3). As confidence intervals make analysis independent of the experimental design, data was analyzed only twice: by considering 3 and 9 replications (Table 2). The full dataset (9 replications) was also used three times in a row, in a third analysis, to roughly estimate the effect of replication number on the size of the confidence interval (Figure 3). This data (27 replications), however, is biased since every replication was repeated three times, being useful only to be compared with the 9 replication analysis in terms of the size of the confidence interval.

If the 95% confidence interval bars of different treatments do not overlap, one can be sure they are statistically different (Hsu, 1996). By comparing graphs with different replication numbers (Figure 3), one will observe the 27 (9+9+9) replication analysis presented confidence intervals smaller than the 9 replication one, but not so small as the observed for the 3 replication analysis. This means that supposing the example was done with real 27 replications, confidence intervals would most probably be similar to the observed for the 9 replication analysis. On the other hand, an ANOVA table in this situation

Table 2. Parameters used to obtain the confidence intervals that are presented in Figure 3 for grain yield observations as a function of herbicide treatments.

3 replications						9 replications						27 (9+9+9) replications (estimate)					
Treat.	Mean	SD	SE	LL	UL	Treat.	Mean	SD	SE	LL	UL	Treat.	Mean	SD	SE	LL	UL
1	573.3	31.5	18.2	537.7	609	1	554.1	184.4	61.5	433.6	674.6	1	554.1	177.2	34.1	487.3	620.9
2	1250.3	31.7	18.3	1214.4	1286.2	2	1005	352.2	117.4	774.9	1235.1	2	1005	338.4	65.1	877.3	1132.7
3	971.3	53.6	30.9	910.7	1032	3	824	226.3	75.4	676.2	971.8	3	824	217.4	41.8	742	906
4	1004.3	46.7	27	951.5	1057.2	4	754	328.8	109.6	539.2	968.8	4	754	315.9	60.8	634.9	873.1
5	692.3	17.6	10.2	672.4	712.3	5	558.7	182.4	60.8	439.5	677.8	5	558.7	175.2	33.7	492.6	624.8
6	1227.7	30.7	17.7	1193	1262.4	6	1021.2	304.4	101.5	822.3	1220.1	6	1021.2	292.5	56.3	910.9	1131.5
7	548.3	30.5	17.6	513.8	582.9	7	424.3	152.2	50.7	324.9	523.8	7	424.3	146.2	28.1	369.2	479.5
8	1151.7	53.5	30.9	1091.1	1212.2	8	973.6	251.6	83.9	809.2	1137.9	8	973.6	241.7	46.5	882.4	1064.7

Treat. = treatments; SD = standard deviation; SE = standard error; LL = lower limit of the 95% confidence interval; UL = upper limit of the 95% confidence interval.

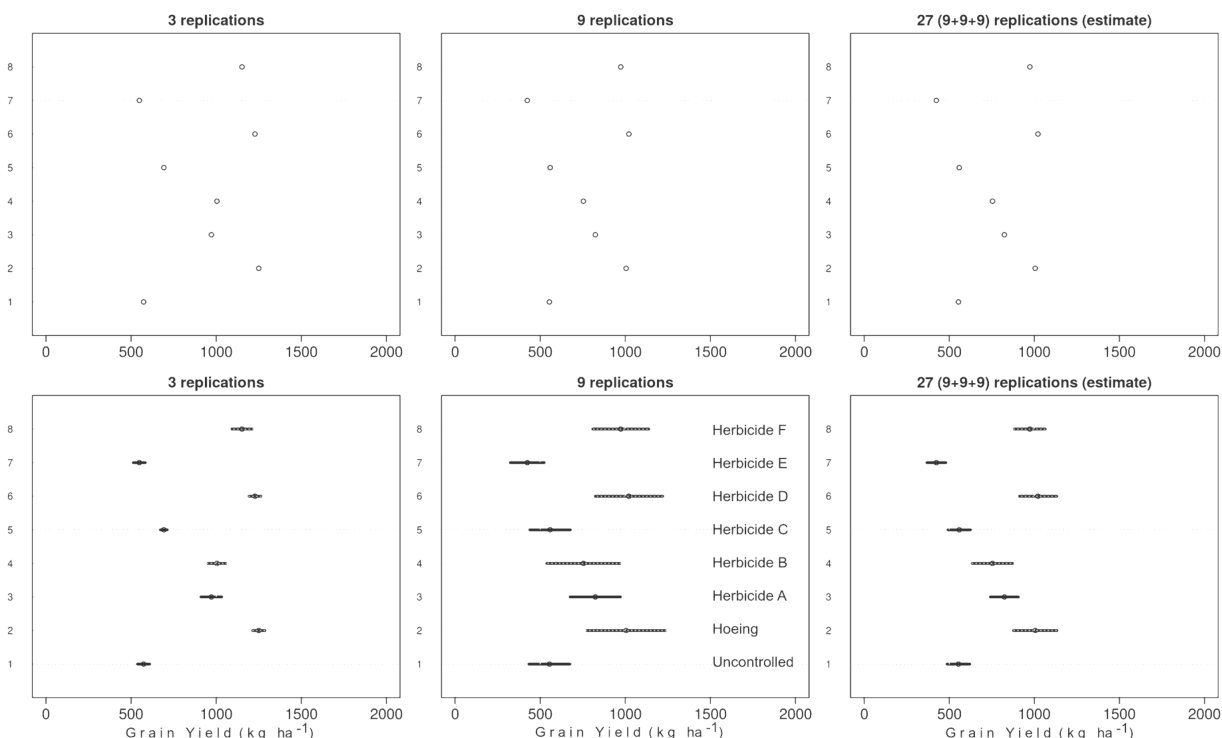


Figure 3. Mean grain yield (Kg ha^{-1}) and 95% confidence interval as a function of herbicide treatments and number of replications.

would be most prone to overestimate treatment effects (Dahiru, 2008), as previously discussed.

By using the confidence intervals for mean comparison, there is the need to adapt the way researchers are familiar to discuss the results; it shifts from the “treatments differed/not differed” to “the difference between these treatments is between X (lower limit) and Y (upper limit)”, since confidence intervals supply effect sizes. Let’s use as example for a brief discussion the 9 replication RBD analysis at 5% probability at Table 1, compared to the 9 replication analysis by 95% confidence intervals (Figure 3), for a very summarized comparison (Table 3).

Results discussion based on confidence intervals are also different because the inferences always refer to what will be observed for all *Jatropha* production fields under equal edapho-climatic conditions (estimation of the real population mean). This does not mean that if we install another identical experiment the new means will

be between the previously established intervals (Steel and Torrie, 1980) because the sample means are always centered on the population mean (Hsu, 1996), but similar intervals most often occur if correctly powered experiments are repeated (Cumming et al., 2007). Because of this, the experiment should have enough power and be carefully planned, installed, conducted and evaluated (Reinhard, 2015). Moreover, the number of replications and the sampled area per plot should be adequate.

The great advantage of confidence intervals over p-value based methods, is that it includes *effect sizes* (Prel et al., 2009) and allows the researcher to inform the farmer what is the interval of answer to be expected by using a given treatment: “[...] the field will produce between the same and “X” kg ha^{-1} more by using a given herbicide compared to the current treatment”. With p-value methods, the information would be: “[...] the experimental field will produced

Table 3. Differences in interpretation of treatment impacts on *Jatropha* seed production between the p-value based analysis and the one based on confidence intervals. A simple discussion is supplied.

(from Table 1)	(from Figure 2)
Duncan's MRT at 5% probability, 9 replications, RBD	95% confidence intervals, any experimental design
Four groups of treatments; herbicides D and F do not differ from the hoeing treatment; in second place comes herbicides A and B; in third place herbicide C which do not differ from the uncontrolled treatments, and herbicide E in fourth place, with low weed control and high toxicity to <i>Jatropha</i> .	Herbicides D and F show a superior response, being similar to the hoeing treatment; herbicide A results in production which may be equal or until 85 kg less grains of <i>Jatropha</i> per hectare; herbicide B may reduce grain yields between zero and 243 kg ha ⁻¹ of grains compared to the lower possible production observed for the best treatments (hoeing in this case); the other herbicides are not of interest because they reduce <i>Jatropha</i> yields too much, being the maximum possible reduction in yields 942 kg ha ⁻¹ of grains when using herbicide E.

“X” kg ha⁻¹ more grains, and is different from the current treatment [...]” - which is the mean of the experiment, and has no consistent connection to what will most often be observed in real fields; no information about the superior and inferior limits of real production levels are possible. P-value based statistics will very often result in *truth inflation* (Reinhard, 2015), as in the finding differential *Euphorbia* responses to light previously discussed.

For those who prefer to stick to the p-value based statistics, the advice is to present the confidence intervals to determine the range of the answers consistent with the data, independently of the significance obtained, as defended by Prel et al. (2009), Chavalarias et al. (2016) and Farland et al. (2016). The pre-test (F-test) enthusiasts may use the confidence interval as well to check overall significance of a variable. If the confidence interval at 5% for the whole data set of the variable includes “zero”, there is no significance ($p > 0.05$); if the confidence interval does not include zero, there is significance and treatments should be compared (Motulsky, 2002; Cumming et al., 2004, 2007).

Although there are claims that samples must be normally distributed for the confidence intervals to be valid (Steel and Torrie, 1980), apparently confidence intervals can be used with distributions

that are not normal — that are highly skewed or in some other way non-normal (Carlberg, 2011).

The quantitative data and the confidence intervals

Not only qualitative data can benefit from the confidence intervals, but also quantitative data (Cumming et al., 2007). Supposing in a given situation researchers face the problem of deciding if two methods of herbicide application influence the rate of herbicide drift as a function of the distance from the point of application, the first step is to fit regressions with their respective equations to both sets of data (Figure 4a). Datasets are “visually” different, but the statistician is advised never to trust the bare data but always to statistically process it prior to any inference (Reinhard, 2015).

The most widely used method to determine if two regressions are distinct, is to compare their β (beta) coefficients, although other methods exist (Karlson et al., 2010). This involves some data transformation and processing (variables standardization). Other researchers prefer a more “psychic” method and run the same regression model for every replication (or block) of a treatment. Later, these β -values are used as “replications” to compare equations by LSD or Tukey's. This seems inefficient and probably full of limitations.

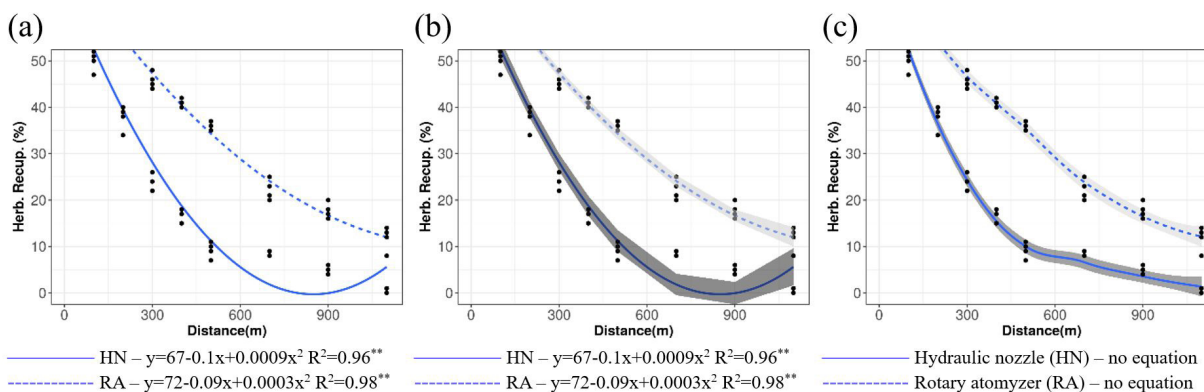


Figure 4. Herbicide drift as a function of distance from the application point with two different spraying methods. Data was analyzed fit to polynomial 2nd degree regressions without (a) and with (b) 95% confidence intervals and to (c) Loess 2nd degree regression with 95% confidence intervals.

Therefore, it is proposed here the use of confidence intervals to infer about regression differences, as in Figure 4b. The existence of the confidence intervals throughout the regression curve allows one to easily determine the data sections where the regressions differ and where they superpose. In the sections where the confidence intervals do not overlap, the regressions are clearly different at the same confidence level (Reinhard, 2015). One will be able to compare in a glance as many regressions as needed by observing their confidence intervals. When applied to regressions, “confidence intervals” may be named “confidence bands” as well (Steel and Torrie, 1980).

Another advantage of the coefficient intervals in regressions is the ability to discard the equations in cases where the x/y relation does not need to be automatically determined (as the case of use in some software or spreadsheet for instantaneous calculations of “y” by changing “x”). This makes room for most optimal methods of curve fitting as the Loess/Lowess (Cleveland and Devlin, 1988), which permits to locally establish a wider class of regressions compared to the parametric functions as polynomials. If one compares the regression curve of “Hydraulic Nozzle” (solid line, Figure 4a, b) it is noticeable the lack of fit in the interval of $600 < x < 900$, even with $R^2 = 0.96$. This is unlikely to happen when fitting regression

curves by methods as the Loess (Cleveland and Devlin, 1988; Hafen, 2010).

The drawback is the nonexistence of the equation which traditional researchers demand to be obtained in any regression analysis, but in fact they are not always used. Most up-to-date statistical environments are enabled to work with local regression models. Loess and Lowess methods are both built up on classical methods as linear and nonlinear least squares regression (Cleveland, 1979). Its main disadvantages are the need for relatively large datasets in order to produce good models, since it relies on the local data structure (Hafen, 2010); and to be as susceptible to outliers as the parametric regressions (Cleveland, 1979).

Conclusions

This review gives some insights on how flawed choices of statistical methods, specially the *old and good* p-value based statistics, can pave the way for mistaken conclusions in properly conducted experiments in weed research. Therefore, it is here proposed that the use of confidence intervals, as a single or complementary approach, could reduce frequent misunderstandings. Furthermore, this study highlights that great part of the results of herbicide efficacy that have been conducted, are probably based on underpowered experiments

and, therefore, are prone to output some type of distorted data.

Thus, if we care to plan and execute an experiment carefully, it seems logical to dedicate the same effort when reporting findings through statistical analysis. This may ensure high standards on weed research that otherwise can easily turn amazing studies into a half dozen of mistaken findings.

Readers are warned not to think that most of the published data related to the weed science is simply “wrong” or biased due to the statistical limitations we present; even conclusions based on weak statistical methods may still be valid. We only should have less reliability on them than we anticipate.

References

- APA – American Psychological Association. **Publication manual of the American Psychological Association**. 6th ed. Washington: APA, 2009. 272p.
- Burns, R.B.; Dobson, C.B. **Experimental psychology: research methods and statistics**. Rotterdam: Springer, 1981, 439p.
- Carlberg, C. **Statistical analysis: Microsoft Excel 2010**. Indianapolis: Que Publishing, 2011. 412p.
- Chavalarias, D.; Wallach, J.D.; Li, A.H.T.; Ioannidis, J.P.A. Evolution of reporting P values in the biomedical literature, 1990-2015. **Journal of the American Medical Association**, v.315, n.11, p.1141-1148, 2016.
- Cleveland, W.S. Robust locally weighted regression and smoothing scatterplots. **Journal of the American Statistical Association**, v.74, n.368, p.829-836, 1979.
- Cleveland, W.S.; Devlin, S.J. Locally weighted regression: an approach to regression analysis by local fitting. **Journal of the American Statistical Association**, v.83, n.403, p.596-610, 1988.
- Cumming, G.; Fidler, F.; Vaux, D.L. Error bars in experimental biology. **The Journal of Cell Biology**, v.177, n.1, p.7-11, 2007.
- Cumming, G.; Williams, J.; Fidler, F. Replication and researchers' understanding of confidence intervals and standard error bars. **Understanding Statistics**, v.3, p.299-311, 2004.
- Dahiru, T.P. Value, a true test of statistical significance? A cautionary note. **Annals of Ibadan Postgraduate Medicine**, v.6, n.1, p.21-26, 2008.
- Donner, A.; Koval, J.J. The effect of imbalance on significance-testing in one-way model ii analysis of variance. **Communications in Statistics**, v.18, n.4, p.1239-1250, 1989.
- Farland, L.V.; Correia, K.F.; Wise, L.A.; Williams, P.L.; Ginsburg, E.S.; Missmer, S.A. P-values and reproductive health: what can clinical researchers learn from the American Statistical Association? **Human Reproduction (Oxford, England)**, v.31, n.11, p.2406-2410, 2016.
- Goldacre, B. **Bad pharma: how drug companies mislead doctors and harm patients**. London: Fourth Estate, 2012. 364p.
- Hafen, R.P. **Local regression models: advancements, applications, and new methods**. 304 f. Thesis (Doctor of Philosophy) - Purdue University, West Lafayette, 2010.
- Hsu, J.C. **Multiple comparisons: theory and methods**. New York: Chapman, 1996. 296p.
- Huff, D. **How to lie with statistics**. London: Penguin books, 1954. 124p.
- Karlson, A.K.B.; Holm, A.; Breen, R. **Comparing regression coefficients between models using logit and probit: a new method**. Aarhus: Aarhus University, Center for Strategic Educational Research, 2010. 41p. (Working Paper Series, CSER WP No.0003).
- Korczynski, P.M.; Logan, J.; Faust, J.E. Mapping monthly distribution of daily light integrals across the contiguous United States. **HortTechnology**, v.12, p.12-16, 2002.
- Motulsky, H. **The link between error bars and statistical significance**. 2002. Available from:

https://egret.psychol.cam.ac.uk/statistics/local_copies_of_sources/Cardinal_and_Aitken_ANOVA/errorbars.htm. Accessed 15 Mar. 2017.

Perry, J.N.; Rothery, P.; Clark, S.J.; Heard, M.S.; Hawes, C. Design, analysis and statistical power of the Farm-Scale Evaluations of genetically modified herbicide-tolerant crops. **Journal of Applied Ecology**, v.40, n.1, p.17-31, 2003.

Peternelli, L.A.; Mello, M.P. **Conhecendo o R: uma visão estatística**. Viçosa: UFV, 2011. 185p.

Planta Daninha. Viçosa: UFV, 2016. v. 34, n. 1-4. ISSN 0100-8358.

Prel, J.B.; Hommel, G.; Rohrig, B.; Blettner, M. Confidence interval or p-value? **Deutsches Ärzteblatt International**, v.106, n.19, p.335-339, 2009.

Rao, S.; Fein, D.; Seidman, L.; Tranel, D. Editorial. **Neuropsychology**, v.22, p.1-2, 2008.

Reinhard, A. **Statistics done wrong: a woefully complete guide**. San Francisco: No Starch Press, 2015. 116p.

Silveira Junior, P.; Machado, A.A.; Zonta, E.P.; Silva, J.B. **Curso de estatística**. Pelotas: UFPel, 1989. v.1, 135p.

Steel, R.G.D.; Torrie, J.H. **Principles and procedures of statistics: a biometrical approach**. 2nd ed. New York: McGraw-Hill, 1980. 633p.

Ulguim, A.R. **Identificação, caracterização morfo-fisiogenética e habilidade competitiva de biótipos de *Euphorbia heterophylla* L. com resistência de nível baixo e suscetível ao glyphosate**. 2016, 146 f. Thesis (Doutorado em Fitossanidade) – Programa de Pós-graduação em Fitossanidade - Faculdade de Agronomia Eliseu Maciel - Universidade Federal de Pelotas, Pelotas, RS, Brazil.